DOCUMENT RESUME

ED 202 134 EA 013 586

AUTHOR Pfau, Richard H.
TITLE Sources of Invalidity When Comparing Classroom Behaviors Across Cultures and Nations.
PUB DATE Mar 81
NOTE 51p.; Paper presented at the Annual Meeting of the Comparative and International Education Society (25th, Tallahassee, FL, March 18-21, 1981).

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Classroom Observation Techniques; *Classroom Research; *Comparative Analysis; *Cross Cultural Studies; Elementary Secondary Education; Foreign Countries; Student Behavior; Teacher Behavior; *Validity
IDENTIFIERS *Category Observation Systems

ABSTRACT

Focusing on the use of category systems in classroom observation, this report summarizes factors that may significantly affect the validity of cross-national and cross-cultural comparisons of classroom behaviors. Category systems measure well-defined behaviors by recording events observed at specific intervals or as they begin and end. Areas of concern related to the use of this strategy in multicultural comparisons include cost, identification of comparable populations, resistance by some teachers to observation, observer effects on classroom behaviors, the selection of the specific behaviors chosen for observation, the classification of significantly different events in the same category, limitations in the number of behaviors measured, and theoretical and cultural biases. The encoding and decoding processes for category systems also raise several concerns, including inter-investigation reliability, observer accuracy, subjectivity, distortions in recording procedures, problems in context and meaning, and failure of personnel to follow project protocol. Those designing future studies should consider the fact that the effects of many of these factors can be controlled or eliminated. Moreover, persons intimately familiar with the classrooms of the cultures to be compared should become an integral part of almost any investigation in which category systems are used. (Author/WD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED202134

SOURCES OF INVALIDITY WHEN COMPARING CLASSROOM BEHAVIORS ACROSS CULTURES AND NATIONS

Richard H. Pfau
IPS International/School of Education
University of Connecticut
Storrs, CT 06268
U.S.A.

Abstract

This paper reviews a number of factors which may affect the validity of cross-national and other cross-cultural comparisons of classroom behaviors measured using category systems. Factors discussed include those associated with (a) the specific behaviors chosen for observation (such as limitations in the number of behaviors measured, the specificity of behaviors observed, and theoretical and cultural bias), (b) observation of representative samples of classroom behaviors (such as problems of sample drop-out, the stability of behaviors observed, and observer effects upon behaviors measured), (c) a number of encoding concerns (such as inter-investigation reliability, observer drift in accuracy, reactivity of observer accuracy checks, and distortions resulting from the recording procedure used), (d) decoding concerns (i.e., the equivalence of behavioral indices, and problems of context and meaning), and (e) failure to follow a project's protocol.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
R. H. Pfau
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the 25th annual meeting
of the Comparative and International Education Society
Tallahassee, Florida, 18-21 March 1981

EA 013 586

Sources of Invalidity when Comparing Classroom Behaviors

Across Cultures and Nations

This review has been written to summarize what is known about factors which may significantly affect the validity of cross-national and other cross-cultural comparisons of classroom behaviors. It focuses, in particular, upon factors associated with the use of the direct and systematic kind of observation instrumentation known as "category systems." Such systems are characterized by their measurement of specific, well-defined categories of behavior by recording events observed either each time they occur, at very frequent intervals (e.g., every 3 or 5 seconds), at longer but quite precise intervals of time (e.g., exactly every 5 minutes), by recording their duration using a stopwatch, or by noting the exact time behaviors of interest begin and end. As has been discussed elsewhere, these low-inferential characteristics of category systems seem to result in a potential for making precise and valid cross-national comparisons of behavior not equaled by other techniques available to comparative scholars.[1] It will be seen, however, that a number of factors may interfere with the validity of comparisons made using these instruments--factors which should be

considered by comparative researchers thinking of using category systems to compare behaviors across cultures, nations, and time.

Meanings of Validity

Validity may be thought of as referring to the appropriateness of inferences made based upon measurement scores. In a comparative sense, validity indicates that we are measuring in each system under consideration what we intended to measure. This includes the requirement that, when measurements are made of classroom behaviors, differences in scores yielded by an instrument reflect actual differences in behavior--not differences in impressions made on different observers.[2]

In general, the validity of classroom measurements seems to depend upon the fulfillment of at least three conditions. First, a representative sample of the behaviors to be measured must be observed. Second, an accurate record of the behaviors observed must be made. And third, the records must be scored and interpreted so as to faithfully reflect similarities and differences in behavior, and recreate elements of the original situations in their proper perspectives.[3] The following factors associated with each of these conditions are reviewed in this report.

A. Classroom Sampling Concerns

1. Population Comparability
2. Expensiveness of Observations
3. Stability of Behaviors Observed
4. Sample Drop-Out
5. Observer Effects Upon Classroom Behaviors

B. Behaviors Chosen for Observation

1. Characteristics of Categories
2. Limitations in the Number of Behaviors Measured
3. Specificity of Behaviors Observed
4. Theoretical and Cultural Bias

C. Encoding Concerns

1. Inter-Investigation Reliability
2. Observer Drift in Accuracy
3. Reactivity of Observer Accuracy Checks
4. Encoding Subjectivity and Complexity
5. Distortions Resulting from the Recording Procedure Used

D. Decoding Concerns

1. Equivalence of Behavioral Indices
2. Problems of Context and Meaning

E. Failure to Follow Project Protocol

These factors were identified during a comparative study of Nepalese and United States classroom behaviors,[4] as well as during subsequent reviews of methodological problems associated with the use of direct observation techniques and by reading reports of other investigations of classroom events conducted around the world. In the discussion which follows, many of the difficulties caused by these factors as well as strategies for dealing with them are illustrated by referring to one

particular category system, the Flanders Interaction Analysis Categories (FIAC).[5] This has been done partly because the FIAC seems to be the most widely used category system in the world, partly because it has been used by this investigator for cross-cultural work and is well known to him, and because it serves to nicely illustrate a number of potential pitfalls which may confront the comparative scholar. The factors discussed, however, are of general concern and importance no matter what category system is used in a comparative study.

Classroom Sampling Concerns

Population Comparability

Researchers wishing to compare classroom behaviors between countries need to judge the comparability of the populations in each country from which samples might be drawn, and measurements made. This may involve considering questions such as the following. How similar does grade level need to be in order to help ensure that meaningful comparisons can be made (as when college starts after 10 years of schooling in one country and after 12 years in another)? Can subjects given the same labels in different countries be compared, and what should be done if labels differ? Can behaviors occurring in traditional classroom settings be meaningfully compared with behaviors in open classroom situations? Can teacher and student behavior occurring in classes containing only two or three students (as are sometimes found in Nepal, for example) be meaningfully compared with behaviors occurring in classes containing 50 to 60 students?

Such questions about comparability are faced by any researcher wishing to study classroom behaviors cross-nationally, no matter what data gathering techniques are being considered for use--be they questionnaires, interviews, anthropological approaches, or category systems. Being common to any comparative investigation and not particularly unique to the use of category systems, this factor of "population comparability" will not be discussed a great deal here, except to point out that (a) judgments of comparability need to be based in part upon one's research purpose (i.e., why one wants to make comparisons), since from some perspectives things may be judged comparable while from others they may not be--and consideration of research purpose provides a particularly relevant perspective for making such judgments;[6] (b) judgments of comparability should presumably also be strongly based on a knowledge of the classrooms and associated cultures to be compared; and that (c) decisions about comparability are but a few of many basically subjective decisions associated with the use of seemingly quite objective category systems--but decisions which can affect the validity of many inferences made during data analysis, as will be discussed more later.

Expensiveness of Observations

Another area of concern more directly related to the characteristics of category systems is the cost of making observations using these systems. Classroom observation recordings are expensive to obtain--generally much more expensive in terms of time and money than data

gathered using relatively indirect observation techniques such as questionnaires. The expensiveness of direct observation techniques tends to place restrictions upon the sizes of samples selected for observation in terms of the numbers of teachers and students observed, the amount of time spent observing, as well as the numbers of different category systems which are used in a study. These in turn affect the representativeness of samples selected as well as the accuracy of measurements which can be made.

A limiting factor when using category systems is therefore their cost. Apparently as a result, past studies of teachers in which group means of behavioral measurements were computed have rarely included observations of more than 30 teachers, with only a relatively few observations made of any one teacher. When representative descriptions of the behaviors of individual teachers or students are needed (as opposed to group means), the total number of observations made and time spent observing individual persons must usually, of necessity, be increased.[7] This is mainly because many behaviors measured using category systems are so unstable that more than 5 or 10 observations of the classes of individual teachers may be required to obtain accurate estimates of their occurrence.[8] The cost of making such extensive observations in the face of a limited budget effectively places limits on the number and hence accuracy of measurements made, and as a result, limits the accuracy of comparisons made between countries and cultures.

Stability of Behaviors Observed

Another factor, alluded to in the previous section, which may affect the representativeness of behaviors observed, and the validity of comparisons later made, concerns cycles of classroom behaviors that may occur over time. As Flanders has pointed out:

> Interaction patterns cycle through the school year, the semester, the 2- to 4- week unit of study, the day, the daily lesson, and in terms of learning to spell a particular word in a spelling lesson. . . . The relationship between these factors and interaction in the classroom is sufficiently strong to lead to incorrect conclusions if they are ignored.[9]

Thus, care should be taken to ensure that differences in phases of instruction do not bias comparisons made between different groups. Unfortunately for researchers wishing to cope with this factor, little empirical information now exists describing the kinds of cyclical differences which occur during classroom instruction around the world.

Sample Dropout

Even if sound sampling strategies are used in an observational study, the possibility exists that some school administrators or teachers will not allow classes to be observed. Flanders, for example, found that at least 10% and sometimes 50% of United States teachers selected for observation, refused to permit their classes to be observed.[10] In the United Kingdom, a study of over 200 teachers in 72 secondary

schools experienced a somewhat lower 9% non-participation rate, while during a study conducted in Nepal of 72 primary and secondary schools, apparently no headmasters or teachers refused to permit observations of teaching to be made.[11] Apparently, as Ryans has pointed out, in a decentralized system of education such as that found in the United States, participation of a school system, school, or teacher in a research project must usually remain voluntary[12]--and such voluntarism may tend to increase the possibility of non-participation. For whatever reason, though, this kind of dropout when it occurs can make a sample observed less representative.

There are no clear-cut solutions to this problem. Substituting a teacher who is willing to cooperate for one who is not willing, by no means ensures representativeness.[13] Also, if rigorous sampling procedures are used, it seems likely that some teachers will be selected who would rather not be observed. These teachers, if they are forced to participate, may alter their behavior more than volunteers. Thus, more rigorous procedures may produce less representative classroom interaction data.[14]

When considered with a cross-national perspective, the influence of nonparticipants upon the validity of comparisons made between countries is complicated by the likelihood that different types of dropout will occur in the culturally different populations being studied.

Observer Effects upon Classroom Behaviors

Another sampling problem is derived from the fact that an observer's presence may affect the classroom behaviors being measured. That is,

under the gaze of an observer, teachers and students may behave differently than they normally would. After all, teaching is usually a fairly private activity, not conducted in the presence of adults who have intruded into the classroom.[15] If these "intruding adults" represent an authority whose actions might affect a teacher's job or prestige, if they bring with them TV cameras and other videotaping equipment (needed, for example, to rigorously analyze classroom nonverbal communication),[16] if they place small radio-microphone transmitters around teacher and student necks to obtain more complete records of talk which occurs, or if they leave time-lapse photographic equipment in the classroom which take pictures at regular intervals,[17] it is easy to understand that teachers as well as students might behave other than normal.

Research concerning this factor seems clear. Observers can have an effect on classroom behaviors. Samph, for example, in a study of 10 US teachers found that teacher verbal behavior varied as a function of an observer's presence or absence in the classroom. Teachers used more praise, more acceptance of student ideas, and less criticism when they knew they were being observed.[18] Other studies also have found changes in teacher and student behaviors while being observed, although not all behaviors changed and, in some cases at least, observer effects on behaviors studied seemed to disappear after several days.[19]

In partial contrast, other classroom researchers in the United States, United Kingdom, and Nepal have indicated that observer effects in classrooms studied appeared to be negligible.[20] Hilsum and Strong, for example, concluded that "while the distortion caused by an observer's presence could never be eliminated completely, it could be reduced almost

to the point of disappearance..." by using techniques such as those to be discussed shortly.[21] It might be noted though that reports of little or no effects, have usually not been based upon systematic measurement of behaviors observed. Rather, they have been based upon general impressions of a researcher or his assistants or upon teacher and student reports of "no effects," taken at face value. In the latter case, such reports of no effects are not an observed fact, but as Gussow astutely points out, "a self-conscious report by a group about itself."[22] Thus, the evidence supporting conclusions of "little or no effects" has not been particularly strong.

At least four factors have been identified which may serve to influence behaviors observed. These are: 1. observer conspicuousness (presumably, the more novel and conspicuous the observer, the more distracting he or she is); 2. personal attributes of the observer (ex., the observer's professional status, socio-economic class, sex, race, friendliness, dominance); 3. The rationale for observation (are teachers being evaluated, will observations be shared with the principal or superintendent,....?); and 4. differences in those observed (different people may react to observation differently--as might teachers of different cultural backgrounds).[23] However, a number of persons feel that these factors can often be controlled, so that observer influence need not be a great problem. Medley and Mitzel, for example, indicated that if teacher and school anonymity are guaranteed, if teachers are ensured that their behaviors are not being evaluated, and if teachers have an idea of why the observations are being made, then observer effects can be minimized.[24] In a similar vein, Kerlinger felt that

Individuals and groups seem to adapt rather quickly

> to an observer's presence and to act as they would usually act. This does not mean that the observer cannot have an effect. If means that if the observer takes care to be unobtrusive and not to give the people observed the feeling that judgments are being made, then the observer as an influential stimulus is mostly nullified.[25]

Thus, to some degree, it appears that observers can have an effect upon classroom behaviors observed. These effects can apparently be minimized by conducting observations in a nonthreatening and unobtrusive manner. Observer effects also seem to diminish over time. However, the behavioral effects of being observed still remain largely unknown.[26]

It should be realized that when observer presence affects classroom behaviors measured in one or more of the countries or cultures being studied, the validity of comparisons and interpretations later made may be reduced. It is important, therefore, that such effects be lessened as much as is possible--especially since the potential for these effects is usually inherently associated with category system usage.

Behaviors Chosen for Observation

Another aspect of sampling is associated with selection of the specific categories of behavior to be observed and measured. Since all behaviors are not measured by a category system, the behaviors expressed in these systems are also the result of a sampling process. Aspects of such sampling which may affect the validity of inferences and comparisons later made include those which follow.

Characteristics of "Categories"

A "category" can be considered as a set of specifications regarding what events will be grouped as equivalent.[27] This is because categories are used to classify and group together similar but differing events (and objects), with distinctions between the events classified being lost during the encoding process. For example, the single category "teacher question," if used, could classify together questions asking for rote responses, questions asking for student opinions, quite short questions, as well as much longer ones. This may lead to difficulties later during the decoding process, if distinctions need to be made between some of the different events which were classified using the same category. Such difficulties may arise, for example, when differing events grouped as equivalent vary in emphasis from one country to another, when the "focal points" or core meanings that are judged to best represent the category vary from culture to culture, or when category boundaries vary among cultures so that an event judged to be classifiable using the category by members of one culture is not classified using that same category by members of another culture.

For example, let us consider the FIAC category "pupil talk-initiation." When this category is used, what would happen during the decoding and analysis process if "pupil initiation" in one nation consisted mainly of students expressing their own ideas and opinions in an extended manner, while in another nation such initiation consisted mostly of students asking questions of the teacher about content matter, or of "off-target"

remarks by students and resistance to compliance. Part of the answer, of course, lies in why one wants to make measurements of student initiation in the first place, for depending upon one's purpose, the validity of comparisons and interpretations later made may or may not be affected. Part of the answer also should lie in the results of preliminary studies conducted in classrooms of the cultures to be compared.

That is, if important differences in behaviors classified using a single category become apparent during preliminary studies such that the validity of inferences later made may be affected, steps can be taken to preserve those differences in recordings made. These steps involve "subscripting" the categories of concern. Subscripting means dividing a single category into additional subcategories.[28] For example, the FIAC category "pupil talk--initiation" which is usually represented by the number 9 could be subscripted to differentiate between some of the behaviors previously mentioned. Thus, the code number 9.0 could be used to indicate questions asked by students of the teacher, 9.1 could be used to indicate student expression of their own ideas and opinions about the subject-matter being discussed, 9.3 could be used to indicate unruly, non-academic remarks by students, and so on (with care being taken so that each sub-category used is adequately defined so that observers can reliably differentiate between them). In this manner, distinctions of momentary or more permanent interest could be provided, while allowing the data gathered to be collapsed back into broader categories if deemed appropriate during later decoding and analysis.[29]

Such subscripting might also be used to help ensure comparability

in the face of differences in behaviors between the cultures being studied. As Armer has indicated, "one common procedure for reducing cultural specificity is to raise the concepts involved to a higher level of abstraction.[30] Thus a supra-category might be generated to subsume similar but different sub-categories of behavior, and subscripting used to differentiate between those sub-categories during the measurement and analysis process. In addition (and even more speculatively since little work has been done in this regard) if certain important category boundaries differ from one culture to another, such that certain events are not considered a part of a particular category in one culture, but are considered classifiable using that category in another culture, the use of a subscripted category would seemingly allow such differences to be taken into account during later data analysis.

Using subscripted categories in these ways may help solve a number of problems which, on first glance, might cause comparativists undue worry. However, the technique will work only up to a point, since categorization inevitably involves at least some simplification of the events classified. Such losses are probably necessary costs of cross-cultural comparability. A judgment that must be made when comparative studies are being considered is whether the losses are justified by the benefits accrued by using such coding devices.[31]

Limitations in the Number of Behaviors Measured

Category systems, by their very nature, focus observation upon relatively limited numbers and specific types of behavior. Typically, each instrument deals with only one particular dimension of classroom

occurrences and uses less than 10 or 20 categories to measure occurrences in that dimension.

Although the numbers of dimensions and categories of behavior measured can be increased by having an observer use more than one category system, by using more than one observer each of whom uses a different system, by subscripting categories, or by using media devices such as videotape recorders to preserve classroom events for later analysis, each of these approaches has limitations. For example, using more than one observer or videotaping equipment may unduly affect the naturalness of behaviors observed and will increase the costs of data gathering. Similarly, training an observer to use more than one category system or to use a subscripted system may increase the length and subsequent costs of training. There are limits also in the numbers of categories that an observer can record, especially when events are rapidly changing.

Such limitations in the number of behaviors actually measured have importance especially during the process of decoding and analyzing the information obtained. More related to this factor will be discussed under "Problems of Context and Meaning."

Specificity of Behaviors Observed

A characteristic of category systems which apparently helps make them especially suitable for comparative purposes is that they usually focus upon relatively specific behaviors. This is fine if one is only interested in measuring and comparing such behaviors. However, additional questions of validity arise if an investigator wishes to use the

direct measurements made with a category system to infer the magnitudes of less concrete variables or constructs such as teacher clarity of presentation, teacher effectiveness, or stages in the development of educational systems such as C. E. Beeby and Wilbur Schramm have proposed.[32] When such inferred measurement statements are made, questions of validity revolve around the appropriateness of the inferences made from measurements of the specific behaviors measured. This is because the validity of inferred measurements and comparisons based on them is not nearly as straightforward as the validity of the more direct measurements made using a category system.[33] Such inferred measurement needs to be cautiously approached.

One implication of this is that investigators interested in measuring and comparing variables not directly measurable by using a category system should consider using other techniques, such as rating systems, to estimate magnitudes of the variables of concern. In cases when category systems are chosen to make inferred measurements, careful consideration needs to be given to both the content validity of the behavioral categories chosen to indicate the variable of concern as well as to the construct validity of categories used for such purposes.

More related to this area of concern is discussed later, under "Equivalence of Behavioral Indices."

Theoretical and Cultural Bias

Lastly, cultural biases associated with the behaviors chosen for measurement seem worth considering. Given that nearly all category systems to date have been either developed in "Western" cultures,

adapted from instruments developed in the West, or based upon theory developed in Western cultures, the probability as well as potential danger of cultural bias exists when these instruments are used.

That categories of behavior chosen for observation nearly always reflect cultural or theoretical bias seems clear. This is because one's theoretical or ideological system of reference tends to orient what one observes.[34] As a result, "any observational system necessarily reflects the values of the inventor of that system, in that it focuses attention on the behaviors that he or she thought were important enough to study."[35]

For example, Walker has pointed out that the Flanders Interaction Analysis Categories (FIAC), mentioned before, clearly reflect American moral commitments to liberal political attitudes and a democratic conception of personality. He stated that, in the United States,

> the implicit concept of the school is one in which teacher-child relationships are crucial in the regeneration of national culture: authoritarianism and domination are seen not only as psychologically damaging to the child but as a cause of serious social problems.[36]

That the FIAC reflects such orientations seems evident, for its categories of "indirect" and "direct" teacher behaviors, and its concern with response and initiation in the classroom are built upon a research tradition which includes H. H. Anderson's studies of the "dominative" and "integrative" behavior of teachers, Lippitt and White's studies of "authoritarian," "democratic," and "laissez-faire" leadership, and Withall's categorization of behavior as "learner-centered"

and "teacher-centered."[37] Flanders' own words indicate this orientation.

> In the Flanders system of interaction analysis observation all teacher statements are classified first as either indirect or direct. This classification gives central attention to the amount of freedom the teacher grants to the student. In a given situation, therefore, a teacher has a choice. He can be direct, minimizing the freedom of the student to respond, or he can be indirect, maximizing the freedom of the student to respond.[38]

The authoritarian/democratic dimension underlying the FIAC thus reflects a "general quality of American culture rather than a real quality of teacher behavior."[39] Such is the case for nearly any dimension or category of behavior chosen for observation--it is at least initially, and probably finally, either ideologically, theoretically, or in some other way, culturally dependent.

However, one way to avoid at least blatant ethnocentric biases could be to adapt or develop new category systems using a "decentering method" similar to that used when establishing translation equivalence.[40] Such a method would entail having people from the different countries and/or cultures to be studied and compared jointly prepare observational instruments which appear equally applicable to the study of behaviors in their differing classroom situations. Such persons would also be involved during the decoding and analysis process, as will be discussed later, to help ensure that meaningful comparisons are made.

Theoretical and cultural biases which occur may, however, not

necessarily be "bad" (in this writer's opinion at least). The use of existing instruments, for example, has the advantage of permitting comparative scholars to more directly build on previous research -- including the making of comparisons with data previously gathered by others. Also, it seems as if final judgment of the worth of comparisons made using existing, "biased" instruments should be dependent upon the empirical results obtained, rather than upon an unwavering insistance that culturally biased instruments are not useful. The choice of behaviors to be observed should depend upon the purposes of a study, and for some purposes, biased choices may be appropriate, or at least tolerable. However, at the least it seems appropriate that comparative researchers attempt to become aware of the orientations of category systems which they may use, attempt to determine when those orientations may affect the meaningful comparability of behaviors measured and make appropriate modifications when needed, and make explicit the vantage point taken so that others can better judge results of the measurements and comparisons made.

Encoding Concerns

Introductory Remarks

In addition to the sampling strategies used, the ways in which classroom behaviors are encoded can have an effect upon the validity of comparisons made later. "Encoding," as used here, refers to the conversion of observed behaviors into a form suitable for counting and tabulation. Encoding occurs when a behavior is observed and an

observer classifies it using a category and records the related code symbol.[41]

The validity of comparisons can be reduced if different observers encode behaviors differently. As will be seen, differences in the ways that behaviors are encoded can result from conscious decisions made by researchers, unconsciously, and as a result of the way that a category system is designed and recordings are made using it.

Inter-Investigation Reliability

A researcher studying the classroom behaviors of different countries may wish to compare existing data gathered by other researchers who used the same category system. Or, he may conduct or coordinate studies aimed at describing classroom behaviors in the countries of interest, and then compare the resultant descriptions. A major consideration of such a researcher should be the "inter-investigation reliability" of the data compared.

Inter-investigation reliability refers to the similarity of usage of the same instrument during different investigations.[42] That is, it refers to the question of whether the same observation instrument was used the same way by different investigators, or by the same investigator during different studies, so that data gathered may be validly compared.

The need for similarity of instrument usage if valid comparisons are to be made was summarized by Przeworski and Teune when discussing direct measurements and comparative research.

> The rules of empirical interpretation of the terms of a language must be applied consistently to all observa-

> tions. . . .To the extent that the rules are not applied consistently, there is a loss in reliability. To the extent that inconsistent application of the rules tends to be concentrated in one country rather than be distributed across all countries, there is a systematic error in observations. Such error must be removed if the findings are to be reliable.[43]

The kinds of inter-investigation reliability problems which may be encountered are illustrated by a number of practices discovered by this investigator during his study of Nepalese and US classroom behaviors. On the one hand, it was found that researchers often did not use the same category system in exactly the same way. For example, it was discovered that some researchers differed in their use of the FIAC by recording teacher writing on the blackboard or overhead projector by using category 5 (lecturing), versus the standard practice of recording all types of silence, including that occurring while the teacher writes on the blackboard, by using category 10 (silence or confusion); by recording all teacher repetition of student answers by using category 3 (accepts or uses ideas of pupils), versus the practice suggested by Amidon and Flanders and used in the Nepal study, of recording teacher repetition of correct student answers by using category 2 (praises or encourages); or by using the FIAC continuously, from the start of a class period to the end, or by recording and analyzing whole class discussions as well as less public talk with individual students which occurred during laboratory and "seat work" situations, versus the practice (used in the Nepal study) of recording only periods of teacher lecture and public discussions which could be heard by the entire class.[44]

In addition, it was found that researchers frequently did not present evidence indicating that they had used a category system in a standardized manner. That is, common practice in most studies reviewed was to determine the extent of agreement between project observers using the same category system. However, only infrequently were efforts made to determine if observers were using the category system in a standardized manner, by seeing if their observations agreed with those made by "expert" users of the instrument. Such accuracy checks with expert standard users are considered important if comparisons are made between data gathered by different investigators, for it is possible that a team of project observers may reach high levels of agreement among themselves when using an instrument, but their usage may differ from that intended by the instrument's developer.

It might be noted that no technique of demonstrated effectiveness seems to now exist which can be used to determine that consistency of category system use is achieved across cultures and nations. Although approaches used by this investigator in his study of Nepalese and US classroom behaviors may be suitable,[45] research is needed to determine the effectiveness of these and other possible approaches aimed at ensuring that category systems are used in standard ways by different observers in different countries.

Observer Drift in Accuracy

During the survey of teaching in Nepal, a phenomenon was noticed called "observer drift," or more technically, "instrument decay." That is, in many cases observer usage of a category system was found to differ

from standardized usage over time.[46]

Other researchers also have encountered this phenomenon. For example, Ryans (1960) indicated that observer accuracy, as measured by correlations of trainee and senior observer assessments at the end of training, was high. However, during his project, this level of agreement was usually not maintained over an extended period of observation, and refresher training was often required.[47] More systematic investigations of such "drift" have also found that when trained observers are separated into groups, those working together often make more similar observations than those working in different groups.[48] The similarities which occur seem to result in part from discussions between co-observers which tend to idiosyncratically modify use of behavioral codes so that more similar recordings are made. Differential usage also seems related to how well behavioral categories are defined--as will be discussed shortly.

The point is that this phenomenon of drift from standardized usage needs to be controlled if maximally reliable (and valid) measurements and comparisons are to be made. Researchers also need to be aware that conducting only one check of observer agreement during an observational study (as is now frequently done) is not sufficient to judge the accuracy of observations made at other times. Readers are referred to procedures used by Roebuck and his colleagues to maintain observer accuracy over a three year period, and to suggestions made by Kazdin for ideas concerning how such drift can be assessed and controlled.[49]

Reactivity of Observer Accuracy Checks

An important argument when justifying the use of category systems

for comparative purposes has been that the use of such systems can be standardized across cultures and nations. The extent to which such standardized usage is achieved has typically been estimated by determining the agreement between measurements made by different observers who used the same instrument while observing the same behavioral events.

It appears, however, that high observer agreement during such accuracy checks does not ensure that an observation instrument is used accurately on other observation occasions. In fact, a growing body of evidence indicates that estimates of observer accuracy obtained under typical conditions of assessment may be highly misleading. In addition to the slow "drift" from standardized usage discussed in the previous section, an immediate and large drop in accuracy can occur with the change from training to actual observation conditions. Also, when spot checks of accuracy are carried out at later times, observer accuracy seems to increase, but drops again during unmonitored observation conditions. That is, observers tend to do a more accurate job of observing during those times when they know that their accuracy is being checked. As a result, estimates of general accuracy as well as recordings of the frequencies of some behaviors may be inflated by 15% to 25%.[50]

The inflated estimates of observer accuracy which occur seem due to reactive effects of the assessment procedures used. They can also occur as a result of using simpler materials during training than during actual data collection, as well as due to a practice used by many investigators of training observers until they score highly on accuracy checks for one or two times in a row, and using these last estimates as

indicators of accuracy during other times (observers thus being judged as accurate if by chance they had a good day or two in a row).[51]

Although the last two practices mentioned above can be controlled, apparently no procedure has yet been demonstrated to prevent the reactive effects of assessment and subsequent serious reduction in observer performance when accuracy is not being checked.[52] However, several procedures have been suggested to reduce or eliminate the seriousness of such drops. These include using random and covert checks of accuracy, assessing accuracy on every observation occasion during a study, and overtraining observers so that when a drop occurs, it would not fall below acceptable levels of accuracy.[53] The use of audio and video tape recordings of classes (which would permit repeated observations by different observers) can also be used to help solve problems of reactivity or at least more accurately determine true levels of observer accuracy--but such recordings, when made, can of course increase "observer effects" upon the behaviors recorded and observed.

At any rate it should be realized that overt assessments of observer accuracy conducted occasionally during a study may not reflect the extent to which observers are using observation instruments in a standardized way during times when their accuracy is not being assessed.

Encoding Subjectivity and Complexity

Encoding observations using category systems is not a completely

objective process. Observers using these instruments constantly make judgments about the behaviors classified. Most of these judgments are guided by relatively clear rules about what should be done. Some judgments, however, are made more subjectively. In the past these have included difficulties in distinguishing between certain behavior categories, as well as aspects of recording such as deciding exactly when to begin and stop recording behavioral events.[54] When such subjectivity results in differences in the ways events are recorded by different observers, it may affect the validity of comparisons made.

Difficulties related to the encoding of observations can be identified and analyzed by empirically determining the likelihood of confusion between categories used in an instrument.[55] Problem areas when found can often be eliminated by providing more explicit definitions of the concerned categories, so that it is clearer to observers when instances of a particular category occur or not. Such clarification of categories and other coding rules can help to increase general levels of observer accuracy as well as reduce observer drift.[56] Unambiguous definitions should also help ensure that persons reading reports of a comparative study will better understand what was actually measured.

Subjective effects may also be confounded by the amount of time available to observers in which to make their judgments and by the ease of discriminating between the different events being classified. For example, at least one study has found that the more categories of behavior used in an instrument, the lower were the accuracy of recordings made. Those results led the investigators of that study to advise that "attempts to obtain finer bits of information should be considered and

weighed against potential reliability loss and the costs and time involved in training observers."[57] Thus, making a category system more complex, as suggested when the idea of "subscripting" was previously discussed, may help to increase validity in some ways but decrease it in other ways.

Distortions Resulting from the Recording Procedure Used

When behavioral categories are selected for observation, a procedure for recording those behaviors also needs to be chosen. Investigators using category systems have a choice, when selecting a procedure, of either recording behavioral events (a) each time they occur, (b) at very frequent intervals, such as every 5 seconds, (c) at longer but quite precise periods of time, such as exactly every 5 minutes, or (d) recording the duration of behaviors observed by using a stopwatch or by noting the times behaviors of interest begin and end.

That these recording procedures are standardized and objective tends to obscure the fact that when they are used, a certain amount of distortion tends to take place during the encoding and decoding process. For example, when recording units based on time are used, distortions may occur when classroom events are occurring at another rhythm--encoding often being different when events break on or between the arbitrary time boundaries being used. Investigators using other kinds of recording procedures such as frequency counts, can obtain different kinds of distortion, because they are often forced to treat all the units counted as if they were the same length.[58] A danger of this is that since the tempo and form of teaching differ from place to place

around the world, different distortions may occur in different countries which could affect the validity of comparisons later made.

For example, the Flanders Interaction Analysis Categories were developed in the United States, and its 3 second recording tempo seems suitable to the pace of behavioral events in most US classrooms. However when used in another society where rapid and short answers are given by students, distortions of the type shown in Table 1 may occur.

Place Table 1 about here

In fact, "interval recording procedures" such as are used in the Flanders instrument have been repeatedly shown to be susceptable to such distortions--especially when behavioral events are recorded which are shorter than the time interval used in the instrument.[59] The occurrence of these distortions has led a number of behavior analysts to suggest that such interval recording methods be considered for use with great caution, and that when selected over other recording procedures, the smallest intervals possible should be used.[60]

Decoding Concerns

Equivalence of Behavioral Indices

A purpose of decoding is to recreate in easily understandable symbols those aspects of the original classroom behaviors which were encoded.[61] This should be done in such a way that interpretations resulting from the decoding process accurately reflect behaviors which took place. If influence of the previously discussed factors is minor, decoding of the recordings made should result in fairly precise

measurement statements and straight forward interpretations of the behaviors encoded. Difficulties can arise, however, when inferences are made based upon those direct measurement statements.

In addition to difficulties associated with the absence, irrelevance, or differing distributions of variables in different cultures,[62] researchers seeking to describe and compare classroom behaviors may encounter problems during data analysis associated with (a) similar but functionally different behaviors, and (b) functionally equivalent but different behaviors.

> Perhaps a variety of instructional procedures exists for attaining the same educational goals. These behaviors may differ in overt properties and thus may be coded differently by an observer, but they may engender the same implicit behaviors in students. . . . For example, teacher probing of student responses might be equivalent to the use of student ideas; homework questions requiring higher level processing might be equivalent to the higher level teacher questions in classroom interactions; requiring students to respond at frequent intervals might be equivalent to class discussion and frequent questions; teacher disapproval which is followed by eliciting student clarification of what the student said might be equivalent to teacher use of student ideas.[63]

The functional equivalence of such behaviors is of concern when inferences are made from measurements of those behaviors. That is, it

is of concern when inferences from the same kinds of direct measurements to inferred measurements are not equally valid for all systems being studied and compared.[64]

Difficulties which may be encountered can be illustrated by examining the applicability to other cultures of inferential practices routinely used in the United States based upon observations made with the FIAC. It may be recalled that when the FIAC is used, teacher statements are considered to represent either "indirect" or "direct" teacher behaviors. Indirect teacher behaviors (as viewed by Flanders) are behaviors which tend to maximize the freedom of students to respond, while direct teacher behaviors are those which tend to minimize the freedom of students to respond. The FIAC categories "accepts feeling," "praises or encourages," "accepts or uses ideas of pupils," and "asks questions" are considered to represent relatively indirect behaviors, and the categories "lecturing," "giving directions," and "criticizing" are considered to represent direct behaviors. Measurements made using these categories are routinely used to calculate the values of several indices whose magnitudes are used to infer the extent of indirect and direct teacher influence.

Although such inferences may be appropriate to US classroom behaviors, problems of equivalence have been discovered when the same procedures are used in other places. That is, in Nepal, this investigator judged that "teacher questioning" may not represent an indirect behavior as much as a direct one, and "giving directions" may often be considered more an indirect behavior--these interpretations being exactly the opposite of those made in the USA![65] Similarly, in Puerto

Rico when analyzing English as a Second Language (ESL) classes, Velazquez felt that short commands such as "you," "again," and "everybody" represented indirect teacher influence, whereas in the USA, such commands or directions are considered to be direct teacher behaviors.[66] As a result, the indices of teacher indirectness and directness usually calculated in the USA may not validly indicate the extent of such influence in Nepal or in Puerto Rican ESL classes, and thus may not be used to make meaningful comparisons between those places.

The detection and possible remediation of such problems of equivalence during the decoding, data analysis, and interpretation process seems another reason why persons who are intimately familiar with the cultures being studied should be involved in comparative studies of the kind being discussed. Without the insights which such persons can provide, naive researchers unfamiliar with the cultures being studied may make the mistake of happily comparing numbers while making grave errors of interpretation.

Problems of Context and Meaning

Observers using a category system record only behaviors described by the particular system being used. When this is done, they and data interpreters essentially end up ignoring much of the meaning of behaviors recorded and the context in which the behaviors took place. In part, the assumption is made that classroom events can be analyzed in terms of small bits of behavior that have the same value irrespective of context.[67]

Disregarding the context of events recorded and the meaning of

those events to participants, however, carries the risk that recordings made may be later misinterpreted when they are analyzed. This risk seems especially great when category systems are used in cross-cultural studies. Ignoring contextual factors, for example, can place a decoder in the position of not knowing whether students were or were not paying attention to the teacher, of not knowing whether the teacher was interacting with a small group of 2 students or a large class of 60, or of not knowing whether an observer using a category system such as the FIAC had observed "silence" or "confusion" (both of which are recorded by the same FIAC category number). Although this information may not be needed in some studies, in other studies a knowledge of such factors may be relevant and necessary.

A number of researchers, for example, have found that a knowledge of context variables has helped to reveal relationships between classroom variables that would otherwise have been masked, helped to increase the number and interpretibility of findings, and helped to resolve apparent discrepancies and failures to replicate previous studies.[68] In addition, many (if not all) types of behavior, such as bodily gestures, are understandable only if the context in which the behavior occurred is considered.[69] Although only a relatively few contextual variables have been systematically studied so far, some of demonstrated importance to the analysis of classroom events include the social class of students, the purpose behind teacher behaviors recorded (e.g., academic instruction or dealing with problem students), grade level, subject matter, sex of students, group size, and the publicness of interactions recorded (e.g., whole class or private discussion).[70]

How does one decide exactly which contextual variables should be taken into account? Unfortunately, no systematic method seems known which will ensure that all context variables of importance for a particular study have been identified.[71] However, a general approach seems clear. In addition to reviews of past research and thinking, preliminary unstructured observations should be made in classrooms of the cultures to be more formally studied in order to help identify seemingly important contextual variables. Once identified, steps can then be taken to either control for more important variables during the sampling process or obtain information about them when behavioral events are encoded. For example, if it were thought desirable after preliminary studies, category systems to be used might be subscripted to provide additional contextual information (such as distinguishing between "silence" and "confusion"), or provisions might be made to use other data gathering techniques to obtain information about contextual factors--such as, perhaps, determining what percentage of students were paying attention (another category system could be used to determine this), to note physical characteristics of the classrooms studied (a checklist might be used),[72] or to determine the teachers' goals or students' perceptions (using interviews). As indicated, an aim of gathering this "additional" information is to acquire sufficient contextual material in order to avoid misinterpretations due to abstracting the events recorded from their settings.[73] In addition, if one wishes to <u>understand</u> the behaviors recorded, then the use of other techniques to acquire contextual information is almost mandatory. This is because category systems, while providing precise and valid descriptions of many classroom events,

by themselves do not provide much of a basis for understanding and explaining those events!"[74] A real understanding of why behavioral differences and similarities occur between countries seems to require that other information also be considered. Although little along these lines has been done so far, the thinking and work of Delamont and others in the United Kingdom serve to illustrate the benefits which can result by using other techniques to complement investigations of behavior made using category systems.[75]

Failure to Follow Project Protocol

The last potential problem to be discussed here is the failure of personnel to follow a project's protocol (i.e., planned methodology).[76] Such a problem may occur either unconsciously or as a result of dishonesty.

For example, in a study reported by Kent and others, mean observer agreement scores calculated by pairs of observers were .73 and .76, for two different observation conditions. However, recomputation of these scores by a principal investigator and an assistant revealed means of .67 and .70 respectively--these being significantly less than the scores computed by the observers themselves.[77] Similarly, O'Leary and Kent found systematic tendencies for agreement scores to be higher when observers were left unsupervised during reliability checks, as well as a consistent tendency toward errors in computation by observers which spuriously inflated levels of agreement.[78]

During the survey of Nepalese teaching conducted by this investigator, a failure to follow the project's protocol was encountered in at

least three separate ways. Several times, research assistants did not or could not follow the project's sampling plan as far as actually visiting schools selected for observation; through a misunderstanding, one team of observers recorded observations for only part of some class periods observed (rather than the entire periods as planned); and one temporary assistant was found to be fudging data during the decoding process.[79]

Such failures to follow project protocol represent yet another factor which can reduce the validity of cross-national comparisons.

Concluding Remarks

As should now be obvious, quite a few factors may affect the validity of comparisons made using category systems. Some of these factors are of concern to comparative scholars no matter what measurement technique is used. These "general" factors include population comparability, sample drop-out, and failure of personnel to follow a project's protocol. Other factors are somewhat more unique to category system usage. These include observer effects upon behaviors observed, the expense of observations made, distortions associated with the particular recording method used, and problems associated with abstracting events measured from their contexts.

Lest readers become unduly discouraged at the many factors discussed, however, it can be remembered that not all will be equally important during any given investigation, a number of them can be controlled or their effects eliminated, and that apparently valid comparisons have been made in the past by investigators with less information than readers

now have about problems that may be encountered. Presumably, future studies which systematically take these factors into account can be even more effective than past studies which have not done so.

If one looks again at the factors discussed, it is interesting to note how many are related to subjective decisions associated with use of this seemingly objective approach to measurement. To help make many of these decisions, it seems clear that persons intimately familiar with classrooms of the cultures to be compared should become an integral part of nearly any investigation in which category systems are used--from preliminary studies and planning on through to data analysis and report writing. Otherwise, as previously indicated, naive researchers using category systems may make the mistake of happily comparing the numbers generated, while making grave errors of interpretation.

Footnotes

[1]Richard H. Pfau, "The Cross-National Comparison of Classroom Behaviors," (paper presented at the annual meeting of the Comparative and International Education Society, Toronto, February 1976 [ERIC Document Reproduction Service No. ED 120 260]), and "The Comparative Study of Classroom Behaviors," Comparative Education Review 24 (October 1980): 400-414.

[2]Standards for Educational & Psychological Tests (Washington, D.C.: American Psychological Association, 1974) p. 25; Adam Przeworski and Henry Teune, The Logic of Comparative Social Inquiry (New York: Wiley-Interscience, 1970), p. 103; Donald M. Medley and Harold E. Mitzel, "Measuring Classroom Behavior by Systematic Observation," in Handbook of Research on Teaching, ed. N.L. Gage (Chicago: Rand McNally, 1963), p. 250.

[3]Medley and Mitzel, p. 250; Ned A. Flanders, Analyzing Teaching Behavior (Reading, Mass.: Addison-Wesley, 1970), p. 87.

[4]Richard H. Pfau, "A Cross-national Comparison of Classroom Behaviors: Based upon a Survey Conducted within Nepal Using the Flanders Interaction Analysis Categories and Caldwell's Activity Categories Instrument" (Ph.D. diss., University of Pittsburgh, 1977).

[5]Flanders, chap. 2; Edmund J. Amidon and Ned A. Flanders, The Role of the Teacher in the Classroom, rev. ed. (St. Paul, Minn.: Association for Productive Teaching, 1971), chap. 2.

[6]Joseph P. Farrell, "The Necessity of Comparisons in the Study of Education: The Salience of Science and the Problem of Comparability," Comparative Education Review 23 (February 1979): 11-13.

[7]Barak Rosenshine, "The Smallest Meaningful Sample of Classroom Transactions," Journal of Research in Science Teaching 10 (1973):221-226.

[8]Richard Shavelson and Nancy Dempsey-Atwood, "Generalizability of Measures of Teaching Behavior," Review of Educational Research 46 (Fall 1976): 553-611; Oded Erlich and Gary Borich, "Measuring Classroom Interactions: How Many Occasions are Required to Measure Them Reliably?" in Classroom Observation Data: Is It Valid? Is It Generablizable? A Compendium of Methodological Papers, Gary Borich and others (Austin: Research and Development Center for Teacher Education, University of Texas, 1977 [ERIC Document Reproduction Service No. ED 137 371]).

[9]Flanders, p. 156.

[10]Flanders, p. 386.

[11]S. Hilsum and C. R. Strong, The Secondary Teacher's Day Windsor, Berks.: NFER Publishing, 1978), p. 17; Pfau, 1977, p. 36.

[12]David G. Ryans, Characteristics of Teachers: Their Description, Comparison, and Appraisal (Washington, D.C.: American Council on Education, 1960; first Indian reprint ed., Delhi: Sterling Publishers, 1969), p. 58.

[13]Flanders, p. 386.

[14]Ned A. Flanders et al., Teacher Influence Patterns and Pupil Achievement in the Second, Fourth, and Sixth Grade Levels vol. 1 (Ann Arbor: University of Michigan, 1969), p. 5.

[15]Michael J. Dunkin and Bruce J. Biddle, The Study of Teaching (New York: Holt, Rinehart and Winston, 1974), p. 58.

[16]Howard A. Smith, "Nonverbal Communication in Teaching," Review of Educational Research 49 (Fall 1979):660.

[17]E.g., as used by Hilsum and Strong, p. 18; Rob Walker and Clem Adelman, Time-Lapse Photography in Recording Classroom Events (London: Center for Science Education, Chelsea College of Science and Technology, 1971 [ERIC Document Reproduction Service No. ED 047 506]); Graham Nuthall and John Church, "Observation Systems Used with Recording Media," International Review of Education, 18 (1972): 491-507.

[18]Thomas Samph, "Observer Effects Upon Teacher Behavior" (Ph.D. diss., University of Michigan, 1968), and "Observer Effects on Teacher Verbal Classroom Behavior," Journal of Educational Psychology 68 (December 1976): 736-41.

[19]Paul R. Surratt, Roger E. Ulrich, and Robert P. Hawkins, "An Elementary Student as a Behavioral Engineer," Journal of Applied Behavior Analysis 2 (Summer 1968): 85-92; Nuthall and Church, p. 493; Hilda Bengtson Hursh, Donald M. Baer, and Trudylee Rowbury, "A Pilot Project to Examine Whether Teachers 'Turn On' Only When Observers are Present" (paper presented at the annual meeting of the American Psychological Association, New Orleans, August 30-September 3, 1974 [ERIC Document Reproduction Service No. ED 107 622]); Michael Mercatoris and W. Edward Craighead, "Effects of Nonparticipant Observation on Teacher and Pupil Classroom Behavior," Journal of Educational Psychology 66 (August 1974): 512-519.

[20]Ruth E. Arrington, "Time Sampling Studies of Child Behavior," Psychological Monographs 51, no. 2 (1939): 89; Ray Merrill Snider, "A Project to Study the Nature of Physics Teaching Using the Flanders

Method of Interaction Analysis" (Ph.D. diss., Cornell University, 1966), pp. 79-82; Raymond S. Adams and Bruce J. Biddle, Realities of Teaching: Explorations with Video Tape (New York: Holt, Rinehart and Winston, 1970), pp. 24-25; Heather Valerie Birrell, "A Comparative Study of the Verbal Behavior of Teachers in Open Education Classrooms in England and the United States" (Ph.D. diss., Syracuse University, 1974), p. 67; Pfau, 1977, p. 233.

[21]Hilsum and Strong (n. 11 above), p. 19.

[22]Zachary Gussow, "The Observer-Observed Relationship as Information About Structure in Small-Group Research," Psychiatry 27 (August 1964): 246.

[23]Stephen M. Johnson and Orin D. Bolstad, "Methodological Issues in Naturalistic Observation: Some Problems and Solutions for Field Research" in Behavior Change: Methodology, Concepts, and Practice, eds. Leo A. Hamerlynck, Lee C. Handy, and Eric J. Mash (Champaign, Illinois: Research Press, 1973), pp. 45-47.

[24]Medley and Mitzel (n. 2 above), pp. 247-248, & 306-307.

[25]Fred N. Kerlinger, Foundations of Behavioral Research, 2d ed. (New York: Holt, Rinehart & Winston, 1973), p. 539.

[26]Beth G. Wildman and Marilyn T. Erickson, "Methodological Problems in Behavioral Observation" in Behavioral Assessment: New Directions in Clinical Psychology, eds. John D. Cone and Robert P. Hawkins (New York: Brunner/Mazel, 1977), p. 271.

[27]Jerome S. Bruner, "On Perceptual Readiness," Psychological Review 64 (1957): 133.

[28]Flanders (n. 3 above), 1970, p. 126.

[29]Ibid, p. 151.

[30]Michael Armer, "Methodological Problems and Possibilities in Comparative Research," in Comparative Social Research: Methodological Problems and Strategies, eds. Michael Armer and Allen D. Grimshaw (New York: Wiley, 1973), p. 55.

[31]Allen W. Johnson, Quantification in Cultural Anthropology (Stanford: Stanford University Press, 1978), p. 136.

[32]C. E. Beeby, The Quality of Education in Developing Countries (Cambridge, Mass.: Harvard University Press, 1966); Wilbur Schramm, Measuring Educational Development Through Classroom Interaction (Washington, D.C.: Agency for International Development, 1969).

[33]See Przeworski and Teune (n. 2 above), pp. 100-110, for an explanation of why this is so.

[34]Christine Blouet-Chapiro and Gilles Ferry, "Implications of Classroom Interaction Analysis," Classroom Interaction Newsletter 7 (May 1972): 75-83; Ian I. Mitroff, The Subjective Side of Science (Amsterdam: Elsevier Scientific Publishing, 1974), pp. 256-257; Michael Quinn Patton, Alternative Evaluation Research Paradigm (Grand Forks, N.D.: University of North Dakota, 1975), pp. 22-23.

[35]Greta Morine, "Interaction Analysis and Inquiry: Teaching Skills Seen Through a Double Screen," Classroom Interaction Newsletter 10 (Fall 1974): 13.

[36]Rob Walker, "The Sociology of Education and Life in School Classrooms," International Review of Education 18 (1972): 36.

[37]E.g., Harold H. Anderson, "The Measurement of Domination and of Socially Integrative Behavior in Teachers' Contacts with Children,"

Child Development 10 (June 1939): 73-89; Ronald Lippitt and Ralph K. White, "The 'Social Climate' of Children's Groups," in Child Behavior and Development, eds. Roger G. Barker, Jacob S. Kounin, and Herbert F. Wright (New York: McGraw-Hill, 1943); John Withall, "The Development of a Technique for the Measurement of Social-Emotional Climate in Classrooms," Journal of Experimental Education 17 (March 1949): 347-361.

[38]Amidon and Flanders (n. 5 above), p. 6.

[39]Walker (n. 36 above), p. 37.

[40]Oswald Werner and Donald T. Campbell," Translating, Working Through Interpreters, and the Problem of Decentering," in A Handbook of Method In Cultural Anthropology, eds. Raoul Naroll and Ronald Cohen (New York: Natural History Press, 1970), pp. 398-420; Richard W. Brislin, "Back-translation for Cross-Cultural Research," Journal of Cross-Cultural Psychology 1 (September 1970): 185-216; Richard W. Brislin, "Comparative Research Methodology: Cross-cultural Studies," International Journal of Psychology 11 (1976): 215-229.

[41]Bruce J. Biddle, "Methods and Concepts in Classroom Research," Review of Educational Research 37 (June 1967): 338; Flanders, 1970, p. 120.

[42]Barak Rosenshine and Norma Furst, "The Use of Direct Observation to Study Teaching," in Second Handbook of Research on Teaching, ed. Robert M. W. Travers (Chicago: Rand McNally, 1973), p. 168.

[43]Przeworski and Teune (n.2 above), p. 98.

[44]Pfau (n. 4 above), p. 223.

[45]See Pfau (n. 4 above), pp. 283-288, for details of the procedure used.

[46]Ibid, pp. 56-60.

[47]Ryans (n. 12 above), p. 93.

[48]K. Daniel O'Leary and Ronald Kent, "Behavior Modification for Social Action: Research Tactics and Problems," in Hamerlynck, Handy, and Mash (n. 23 above), pp. 87-92; Ronald N. Kent and others,"Observer Reliability as a Function of Circumstances of Assessment," Journal of Applied Behavior Analysis 10 (Summer 1977): 317-324; C. L. Kugle, "A Potential Source of Bias in Classroom Observation Systems: Coder Drift" (paper presented at the annual meeting of the American Educational Research Association, Toronto, March 27-31, 1978 [ERIC Document Reproduction Service No. ED159 203]); Paul T. Marston, Linda K. Zimmerer, and W. Curtis Vaughan, "Coder Drift: A Reliability Problem for Teacher Observations" (paper presented at the annual meeting of the American Educational Research Association, Toronto, March 27-31, 1978 [ERIC Document Reproduction Service No. ED 161 897]).

[49]Flora N. Roebuck and others, Maintaining Reliability in a Longitudinal Study (Monroe, Louisiana: National Consortium for Humanizing Education, Northeast Louisiana University, 1974 [ERIC Document Reproduction Service No. ED 106 730]); Alan E. Kazdin, "Artifact, Bias, and Complexity of Assessment: The ABCs of Reliability," Journal of Applied Behavior Analysis 10 (Spring 1977): 141-150.

[50]John B. Reid, "Reliability Assessment of Observation Data: A Possible Methodological Problem," Child Development 41 (December 1970): 1143-1150; Paul S. Taplin and John B. Reid, "Effects of Instructional Set and Experimenter Influence on Observer Reliability," Child Development 44 (September 1973): 547-554; Raymond G. Romanczyk and others,

"Measuring the Reliability of Observational Data: A Reactive Process," Journal of Applied Behavior Analysis 6 (Spring 1973): 175-184; Ronald N. Kent and others (n. 48 above).

[51]Taplin and Reid (n. 50 above), p. 552; Johnson and Bolstad (n. 23 above), p. 18.

[52]Beth G. Wildman and Marilyn T. Erickson, "Methodological Problems in Behavioral Observation," in Behavioral Assessment: New Directions in Clinical Psychology, eds. John D. Cone and Robert P. Hawkins (New York: Brunner/Mazel, 1977), p. 260.

[53]Johnson and Bolstad (n. 23 above), pp. 20, 22, and 26; Taplin and Reid (n. 50 above), p. 553; Wildman and Erickson (n. 52 above), p. 272.

[54]E.g., see Herta A. Guttman and others, "Reliability of Coding Affective Communication in Family Therapy Sessions," Journal of Consulting and Clinical Psychology 37 (December 1971): 397; Pfau (n. 4 above), pp. 238-241.

[55]Jane A. Stallings and Philip A. Giesen, The Study of Reliability in Observational Data (Bloomington, Ind.: Phi Delta Kappa, 1977 [ERIC Document Reproduction Service No. ED 148 893]); Robert P. Hawkins and Robert W. Dobes, "Behavioral Definitions in Applied Behavior Analysis: Explicit or Implicit," in New Developments in Behavioral Research: Theory, Method, and Application, eds. Barbara C. Etzel, Judith M. LeBlanc, and Donald M. Baer (Hillsdale, N.J.: Lawrence Erlbaum Assoc., 1977), pp. 167-188; Kugle (n. 48 above), pp. 13 and 18.

[56]Guttman et al. (n. 54 above), p. 400; Kent et al. (n. 50 above), p. 323.

[57]Eric J. Mash and John D. McElwee, "Situational Effects on Observer Accuracy: Behavioral Predictability, Prior Experience, and Complexity of Coding Categories," Child Development 45 (June 1974): 375.

[58]Biddle (n. 41 above), p. 341; Rosenshine and Furst (n. 42 above), p. 136.

[59]E.g., Christopher R. Milar and Robert P. Hawkins, "Distorted Results from the Use of Interval Recording Procedures," in Behavior Analysis in Education: Self Control and Reading, eds. Thomas A. Brigham and others (Dubuque, Ia.: Kendall-Hunt, 1976), pp. 261-273; J. Powell et al., "Taking a Closer Look: Time Sampling and Measurement Error," Journal of Applied Behavior Analysis 10 (Summer 1977): 325-332.

[60]Jeanne Altmann, "Observational Study of Behavior: Sampling Methods," Behaviour 49 (1974): 258; Milar and Hawkins, p. 272; Powell et al., p. 331.

[61]Flanders (n. 3 above), p. 120.

[62]Armer (n. 30 above), p. 70.

[63]Rosenshine and Furst (n. 42 above), pp. 172-173.

[64]Przeworski and Teune (n. 2 above), p. 104.

[65]See Pfau (n. 4 above), pp. 252-254 for details.

[66]Clara Velazquez, Teacher-Student Interaction in the Puerto Rican School System (Arlington, Va.: ERIC Document Reproduction Service, ED 142 602, 1977), p. 17.

[67]Rob Walker, "Some General Problems that Arise When Interaction Analysis is Used to Assess the Impact of Educational Innovation," Classroom Interaction Newsletter 7 (May 1972): 44.

[68]Jere E. Brophy and Carolyn M. Evertson, "Context Variables in Teaching," Educational Psychologist 12 (1978); 310-316; Hermine H.

Marshall and Judith L. Green, "Context Variables and Purpose in the Use of Verbal Interaction Analysis," Journal of Classroom Interaction 14 (Summer 1979): 24-29.

[69]Harvey M. Taylor, Americans and Japanese Nonverbal Communication (Ann Arbor, Mich.: English Language Institute, University of Michigan, 1975 [ERIC Document Reproduction Service No. ED 119 489]); Drid Williams, "The Uses of Labanotation in Cross-Cultural Research" (paper presented at the annual meeting of the Society for Cross-Cultural Research, Philadelphia, February 1980).

[70]Brophy and Evertson, 1978; Marshall and Green, 1979; Jere E. Brophy and others, "Grade Level and Sex of Student as Context Variables in Elementary School," Journal of Classroom Interaction 14 (Summer 1979): 11-17.

[71]Brophy and Evertson (n. 68 above), p. 316.

[72]See for example the combination checklist and category system used to describe over 90 aspects of US classrooms, in Jane A. Stallings, "The Development of the Contextual Observation System" (paper presented at the annual meeting of the American Educational Research Association, Toronto, March 27-31, 1978 [ERIC Document Reproduction Service No. ED 166 211]).

[73]Douglass R. Price-Williams, Explorations in Cross-Cultural Psychology (San Francisco: Chandler & Sharp, 1975), p. 18.

[74]Ian Westbury, "Looking in Classrooms: What is Important?," University of Illinois at Urbana-Champaign, 1976 (ERIC Document Reproduction Service No. ED 141 960), p. 3; Sara Delamont and David Hamilton, "Classroom Research: A Critique and a New Approach," in

Explorations in Classroom Observation, eds. Michael Stubbs and Sara Delamont (London: John Wiley, 1976), p. 15; Sara Delamont, "Beyond Flanders' Fields: The Relationship of Subject Matter and Individuality to Classroom Style," in Stubbs and Delamont, p. 109.

[75]Stubbs and Delamont (n. 74 above), 1976.

[76]Theodore Xenophon Barber, "Pitfalls in Research: Nine Investigator and Experimenter Effects," in Travers (n. 42 above), pp. 389-390.

[77]Ronald N. Kent and others, "Expectation Biases in Observational Evaluation of Therapeutic Change," Journal of Consulting and Clinical Psychology 42 (December 1974): 774-780.

[78]O'Leary and Kent (n. 48 above).

[79]Pfau (n. 4 above), pp. 259-260.

TABLE 1

TIME INTERVAL DISTORTIONS OCCURRING IF TEACHER PRAISE AND STUDENT TALK--RESPONSE LASTED FOR ONLY 1 SECOND AT A TIME

Variable	Calculated Value[a]	"Actual" Value[b]
FIAC category		
1. Accepts feelings	.0%	.0%
2. Praises or encourages. . .	6.3%	2.4%
3. Accepts or uses ideas of students.	.0%	.0%
4. Asks questions	13.9%	16.0%
5. Lecturing.	49.5%	56.9%
6. Giving directions.	2.9%	3.3%
7. Criticizing or justifying authority.	1.0%	1.1%
8. Student talk--response . .	13.2%	5.1%
9. Student talk--initiation .	2.3%	2.6%
10. Silence or confusion . . .	11.0%	12.7%
Total talk		
Teacher talk	73.5%	79.7%
Student talk	15.4%	7.6%

TABLE 1 (Continued)

Note. The calculations shown are based upon recordings made in the fifth grade mathematics classes of one Nepalese school.

(a)Based upon standard FIAC calculating procedures, which give equal weightage to each category number recorded.

(b)These idealized values are based on the assumption that teacher praise and student talk--response occurred for only 1 second at a time, whereas other categories of behavior recorded occurred for exactly 3 seconds at a time, or for exact multiples of 3 seconds.